A Novel Method for the Efficient Retrieval of Similar Multiparameter Physiologic Time Series Using Wavelet-Based Symbolic Representations.

Mohammed Saeed[1,2,3] Roger Mark, MD PhD[1,2]
[1]Harvard-MIT Division of Health Sciences and Technology and [2]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA
[3]Philips Medical Systems, Andover, MA

Abstract

An important challenge in data mining is in identifying "similar" temporal patterns that may illuminate hidden information in a database of time series. We are actively engaged in the development of a temporal database of several thousand ICU patient records that contains time-varying physiologic measurements recorded over each patient's ICU stay. The discovery of multiparameter temporal patterns that are predictive of physiologic instability may aid clinicians in optimizing care for critically-ill patients.

In this paper, we introduce a novel temporal similarity metric based on a transformation of time series data into an intuitive symbolic representation. The symbolic transform is based on a wavelet decomposition to characterize time series dynamics at multiple time scales. The symbolic transformation allows us to utilize classical information retrieval algorithms based on a vector-space model. Our algorithm is capable of assessing the similarity between multi-dimensional time series and is computationally efficient.

We utilized our algorithm to identify similar physiologic patterns in hemodynamic time series from ICU patients. The similarities between different patient time series may have meaningful physiologic interpretations in the detection of impending hemodynamic deterioration, and may be of potential use in clinical decision-support systems. As a generalized time series similarity metric, the algorithms that are described have applications in several other domains as well.

Introduction

The ongoing advances in computer processing power, networking, and data storage have enabled modern computers with capabilities of generating, processing and storing terabytes of data. Often the data are time-varying, such as data from biomedical sensors monitoring patients in a hospital intensive care unit (ICU). Massive volumes of data can be readily archived in a digital data warehouse to support research in automated data mining. Ongoing research in time series data mining includes developing algorithms that identify "similar" temporal patterns in a collection of time series [1]. For example, one may consider the following query:

Q1: Identify a group of ICU patients with similar changes in their heart rate and blood pressure trends prior to an episode of severe hypotension.

We have developed a large temporal database of ICU patient records called MIMIC-II (Multiparameter Intelligent Monitoring in Intensive Care) [5]. An ICU patient record can be of varying lengths --- from a few hours to several hundreds of hours of real-world (noisy) physiologic data. Each typical patient record consists of several different data streams that convey unique and important clinical information. There are high resolution (125 Hz) physiologic waveforms of ECG that monitor the heart's electrical activity. Vital signs data (such as blood pressure and oxygen saturation) are acquired at 1 sample per minute. Other clinical data (such as fluid balance and medication drip rates) may be charted on a near-hourly basis.

The MIMIC-II database is a new resource for developing and evaluating temporal similarity metrics. Ideally, a similarity metric should be capable of comparing two records that are of different lengths and consist of multiple

physiologic time series. For example, a slowly increasing heart rate trend over several hours accompanied by a concomitant decrease in blood pressure may be indicative of an internal bleed in a patient. Thus, a similarity metric that fails to capture dynamical relationships between two or more parameters would be of limited use in identifying internal bleeds in a large-scale ICU patient database.

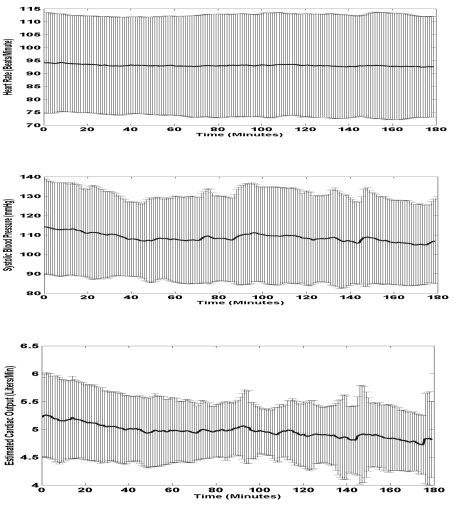

Figure 1: **Multiparameter hemodynamic trends from ICU patients prior to instability.**

Figure 1 includes examples of three channels of physiologic time series (heart rate, systolic blood pressure, and estimated cardiac output) from 88 ICU patient episodes prior to hemodynamic deterioration. The trends are population averaged (with means and standard errors) and demonstrate that there is significant variability from patient to patient. However, data mining may reveal that there are "signatures" in multidimensional physiologic space that may be predictive of hemodynamic deterioration. A number of clinical studies have shown that certain patterns of physiologic deterioration may precede cardiopulmonary arrests [8]. Furthermore, recent studies have suggested that the timely response by clinicians to hemodynamic deterioration in septic ICU patients was the critical determinant of patient survival [7]. Thus, the development of monitoring algorithms that can predict impending deterioration in ICU patients may improve clinical vigilance in a busy ICU.

The paper is organized in the following manner: in the next section, we provide a brief overview of some popular methods used to assess the similarity between time series. Then, we include a methodology for implementing our new trend similarity algorithm. We assess our algorithm's performance using multidimensional time series from real ICU datasets. Finally, we provide a discussion of the major results presented in this paper and suggest possible extensions of our work.

Section II: Review of Previous Methods

Keogh et al [1] provide an excellent survey of methods developed for the retrieval of similar time series. Previous algorithms can be grouped into time-domain methods and transform-based methods. The simplest time-domain algorithm for computing a similarity metric between time series is the Euclidean distance between two discrete time series ***x[n]*** and ***y[n]*** where the distance between the two series is defined as:

$$D(x, y) = \sqrt{\sum_{n=0}^{M} (x[n] - y[n])^2}$$

While the Euclidean distance metric is rather simple, its shortcomings exemplify the challenges in developing more robust time series similarity metrics. The Euclidean distance metric assumes that discrete time series in a database have the same length and are uniformly sampled from their original continuous time processes. In time series recorded in clinical environments, these assumptions are rarely satisfied. To overcome these constraints, modifications to the Euclidean distance metric have been utilized based on the principle of time-warping where signals are "stretched" or "compressed" so that their lengths are the same [1]. However, such signal processing methods may significantly change the unique characteristics of a signal and require careful tuning parameters. Euclidean distance algorithms in particular, and most time-series similarity metrics in general assume that signal are aligned so that "similar" signals will have similar dynamics at the same points in time. Windowing and segmentation techniques have been developed to divide a signal into a set of subsequences which allows greater flexibility in matching time series by using shifting operations [2].

The transform-based techniques project time series of interest onto a set of functions such as sinusoids or principal components [2]. The data transformation reduces the dimensionality of the original times series and facilitates the use of machine learning techniques in matching similar time series. While an improvement over time-domain techniques, transform-based similarity metrics are still an active area of research. Recent attention has also focused on developing symbolic representations of time series [6]. In particular, computationally efficient algorithms for real-time (online) applications are sought that can identify similar multidimensional time series from large databases.

In this paper, we introduce a new temporal similarity metric based on transformation of time series data into an intuitive symbolic representation. This symbolic transform allows us to model a temporal record in a manner similar to popular information-retrieval (IR) models of documents or web pages. The classical IR algorithms utilize a high-dimensional vector-space model in which each element of the vector represents the number of occurrences of a given word in the document [3]. Thus, the structure of the document is characterized by this "term frequency vector" (TFV). In order to transform time-series into a collection of "words" or symbols, a transform is needed that compactly represents the salient characteristics of single and multiparameter time series. We demonstrate that a wavelet-based representation of temporal records offers an intuitively appealing and computationally efficient solution.

In the next section, we describe the methodology used to derive a new wavelet-based symbolic transform and similarity metric.

Section III: Methodology

Wavelets have become increasingly important in areas of signal processing such as data compression, signal de-noising, and feature extraction in pattern recognition [4]. Wavelets are basis functions that can be used to decompose time-varying signals into terms of averages and differences at several different time scales. Several researchers have discussed many of the attractive properties of wavelets [4]. Wavelets have some properties comparable to other popular transform techniques like Fourier analysis. However, wavelets are localized in time, whereas Fourier coefficients represent signal energy components defined over a signal's entire support (time-span). In the present research, we take advantage of the time localization property of wavelets to develop a novel time series similarity metric.

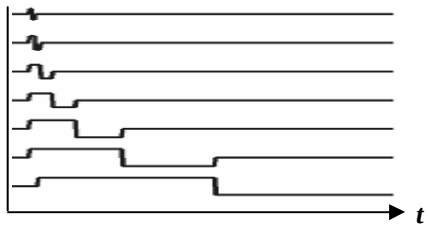

Figure 2: Harr Basis Functions:
By projecting a signal, *x[n]*, on each wavelet basis function, the resulting coefficient characterizes the signal dynamics at the respective wavelet scale. Thus, the coarser wavelet scales can be used to represent changes in a time-series occurring over longer time intervals of several hours, whereas the finer scales can be used to represent changes over a few minutes.

There are several different basis functions that can be utilized in performing the Discrete Wavelet Transform (DWT). Several reviews in the literature provide excellent overviews of the properties of wavelets and can be consulted [4]. The Haar wavelet (see Figure 2) is perhaps the simplest wavelet to implement and can be represented by successive local-differencing and local-averaging operations.

In order to transform a time series into features that can be utilized in classic information-retrieval vector space models, the wavelet coefficients at each scale are quantized into discrete symbols as illustrated in Figure 3. The quantization method that was chosen is the simple uniform quantizer. The number of symbols for each scale is a tunable parameter.

To generate the high-dimensional wavelet symbol vector for a given time series, the elements of the feature vector are assigned values equaling the number of wavelet coefficients (at the respective scale, ***j***) that are quantized to the corresponding wavelet symbol of those elements. Thus, the wavelet symbol vector is similar to a histogram of the frequency of wavelet symbols over all the scales. The wavelet symbol vectors of each scale, ***j***, are concatenated to create one final high-dimensional vector.

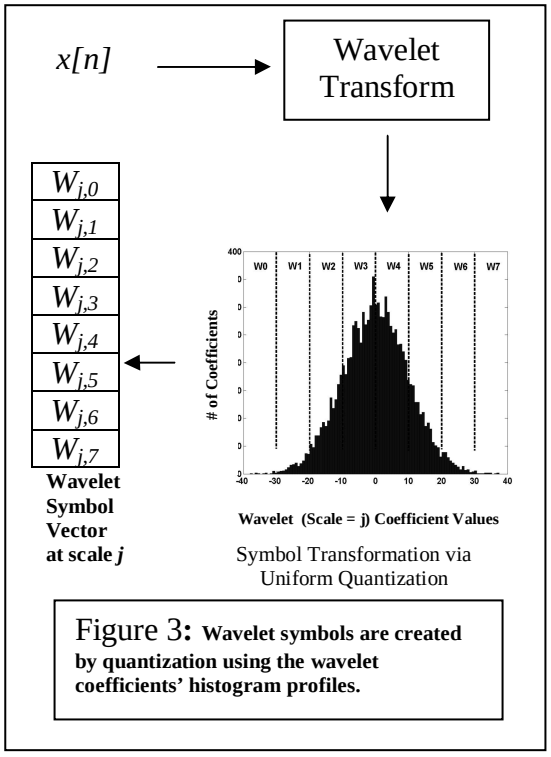

Figure 3: **Wavelet symbols are created by quantization using the wavelet coefficients' histogram profiles.**

The value of each element is then further modified using the popular "Inverse Document Frequency" (IDF) weighting scheme [3]. The IDF weight of an element, $w_{j,i}$, is defined as:

$$IDF(w_{j,i}) = \log \frac{N}{w_{j,i,N}}$$

where N is equal to the number of records in the time series database, and $w_{j,i,N}$ is equal to the number of records in the database that have at least one wavelet coefficient that is quantized to the symbol, $w_{j,i}$. For example, $w_{j,i}$ may represent the wavelet symbol indicating a systolic blood pressure time series, $x[n]$, has a segment of data where systolic blood pressure decreased by approximately 20 mmHg when averaged over a 4-hour time scale. An intuitively appealing aspect of the IDF weighting scheme is that the wavelet symbols that are less frequently observed in the database of time series are more heavily weighted. Thus, the data-driven IDF weighting scheme favors matching time series records that have similar "rare" temporal dynamics.

The final term frequency vector that includes all the wavelet symbols, ($w_{j,i}...w_{J,I}$), of a time series, $x[n]$, is defined as:

$$TFV(x) = [IDF(w_{j,i}) * w_{j,i} ... IDF(w_{J,I}) * w_{J,I}]$$

The distance between the term frequency vectors of two time series, $x[n]$ and $y[n]$ is calculated by computing the correlation coefficient of the two vectors:

$$D(x, y) = p\langle TFV(x), TFV(y) \rangle$$

Section IV: Results

We utilized physiologic trends from a real ICU patient database (MIMIC-II) to evaluate the performance of our wavelet-based similarity metric. Segments of ICU patient records (see Figure 1) including multiple physiologic measurements (heart rate, blood pressure, estimated cardiac output) were selected and categorized into two classes: "hemodynamically stable" and "hemodynamic deterioration." Standard ICU clinical practice includes the use of vaso-active medications in response to a patient's hemodynamic deterioration (sudden and significant drop in arterial blood pressure). Thus, episodes that were labeled as "hemodynamic deterioration" were identified by the administration of a vaso-active medication at the end of the episode. The fiducial point to mark hemodynamic deterioration was selected based upon the time that vaso-active medications ("pressors") were started or significantly increased in patients. For each hemodynamic deterioration episode, segments (sampled at 1 sample/minute) of several hours of data including up to two hours before the fiducial point were chosen to assess the similarity metric algorithm. Thus, the two hours of data prior to the point of deterioration were not made available to the predictive algorithm. The "hemodynamically stable" class included episodes of several hours of physiologic data during which a patient received no significant vaso-active medications or therapies indicative of hemodynamic deterioration.

We hypothesize that hemodynamically unstable patients may have similar physiologic temporal patterns prior to severe decompensation. Using the similarity metric that has been defined, we utilized a K nearest-neighbor algorithm to create a predictor of impending hemodynamic deterioration. The predictor assesses the similarity of a given

multiparameter episode of an ICU test patient to the ***K*** nearest-neighbors from both classes. If the episode to be classified is found to be significantly more similar (based on a threshold, ***t_sim***) to the ***K*** most similar episodes from the "hemodynamic deterioration" class in comparison to the ***K*** most similar episodes from the "hemodynamically stable" class, then the predictor would classify the test patient as being likely to experience hemodynamic deterioration within two hours. The parameters, ***t_sim*** and ***K*** were tuned to optimize the sensitivity and positive predictivity of the classifier. The performance of the predictive algorithm is included in Figure 4. Similarity values greater than 0 were interpreted as increased likelihoods of future hemodynamic deterioration (within two hours). The training data (for the two-class library) consisted of 200 events (100 stable, 100 unstable) from the MIMIC-II database. The test cases included 88 unstable episodes and 89 unstable episodes.

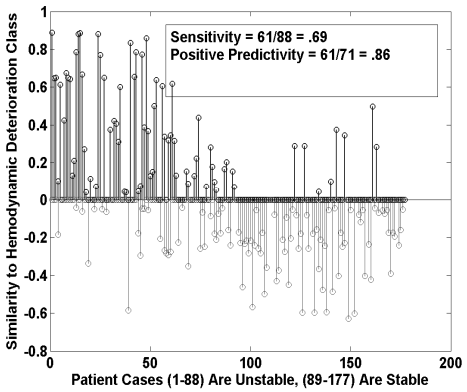

Figure 4: **Similarity (Likelihood) of Hemodynamic Deterioration in ICU Patients**

Section V: Discussion

The major goal of the present paper was to present a new method for assessing similarities between multiparameter physiologic time series. We introduced a novel wavelet-based symbolic transformation that allows for the use of information retrieval algorithms that are popular in the document indexing research community.

Another goal of this work was to investigate if the similarity metric was useful in identifying physiologic patterns that may be predictive of hemodynamic deterioration in ICU patients. The classifier that was developed from the new similarity metric had a relatively high positive predictivity of 0.86. The sensitivity (0.69) can perhaps be further improved by increasing the number of physiologic signals used to assess similarity between ICU patient records. The difficulty in attaining a higher sensitivity is also due to the predictive nature of the classifier. To our knowledge, there is little research in developing automated algorithms that can forecast a blood-pressure drop hours before it happens. The framework presented here can accommodate additional physiologic and clinical signals that can perhaps improve the sensitivity. For example, fluid-balance time series as well as clinical laboratory measurements may further aid in identifying patterns of impending hemodynamic deterioration.

Acknowledgement

This publication was made possible by Grant Number R01 EB001659 from the National Institute of Biomedical Imaging and Bioengineering.

References

1. Keogh, E., Chakrabarti, K., Pazzani, M. & Mehrotra, S. (2001). Locally adaptive dimensionality reduction for indexing large time series databases. In proceedings of ACM SIGMOD Conference on Management of Data, May. pp 151-162.

2. Hetland, M. L., "A survey of recent methods for efficient retrieval of similar time sequences", In Mark Last, Abraham Kandel, and Horst Bunke, editors, Data Mining in Time Series Databases, World Scientific, 2004.

3. J. Rennie, T. Jaakkola. "Using term informativeness for named entity detection." In Proceedings of the 28th Annual Conference on Research and Development in Information Retrieval (SIGIR), 2005.

4. Pietro Lio, "Wavelets in bioinformatics and computational biology: state of art and perspectives," Bioinformatics Vol. 19 no. 1 2003. pp 2-9

5. Saeed et al, "MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring," Computers in Cardiology. 2002;29:641-4.

6. Vasileios Megalooikonomou, Qiang Wang, Guo Li, Christos Faloutsos A Multiresolution Symbolic Representation of Time Series ICDE 2005, Tokyo, Japan.

7. Kumar et al. "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," Critical Care Medicine. 34(6):1589-1596, June 2006.

8. Sanjay et al. "Impact of patient monitoring on the diurnal pattern of medical emergency team activation," Critical Care Medicine. 34(6):1700-1706, June 2006.